

# Mathematical Properties of the Deep Coalescence Cost

Cuong V. Than and Noah A. Rosenberg

**Abstract**—In the minimizing-deep-coalescences (MDC) approach for species tree inference, a tree that has the minimal deep coalescence cost for reconciling a collection of gene trees is taken as an estimate of the species tree topology. The MDC method possesses the desirable Pareto property, and in practice it is quite accurate and computationally efficient. Here, in order to better understand the MDC method, we investigate some properties of the deep coalescence cost. We prove that the unit neighborhood of either a rooted species tree or a rooted gene tree under the deep coalescence cost is exactly the same as the tree’s unit neighborhood under the rooted nearest-neighbor interchange (NNI) distance. Next, for a fixed species tree, we obtain the maximum deep coalescence cost across all gene trees as well as the number of gene trees that achieve the maximum cost. We also study corresponding problems for a fixed gene tree.

**Index Terms**—Deep coalescence, gene tree reconciliation, incomplete lineage sorting, maximal subtrees, nearest-neighbor interchange

## 1 INTRODUCTION

THE minimizing-deep-coalescences (MDC) approach is a promising method for inferring species trees from a collection of gene trees whose discordance is caused by incomplete lineage sorting [1], [2], [3], [4], [5], [6]. In this approach, the amount of discordance for a gene tree and species tree is measured by the deep coalescence cost, which is computed as the total across all edges of the species tree of the number of “extra” lineages required to reconcile the gene tree within the species tree [1]. For a given collection of gene tree topologies, the MDC method identifies a tree that produces the minimal sum of deep coalescence costs over all of the input gene trees. This tree is then used as an estimate of the species tree topology.

The MDC approach has demonstrated favorable performance in several analyses with both empirical and simulated data sets [2], [3]. Lin et al. showed that the MDC criterion has the desirable Pareto property, meaning that a cluster (i.e., the leaf set of a subtree) that appears in every input gene tree must also appear in the MDC-optimal tree [5]. However, we have recently found that under the multispecies coalescent model [8], the MDC criterion is not statistically consistent for asymmetric four-leaf species tree topologies or for species tree topologies with at least five leaves [9]. That is, for some sets of species tree branch lengths, the MDC criterion does not identify the correct species tree topology when gene trees are sampled with probabilities taken directly from the model. It has also been observed informally that the MDC criterion tends to

produce “balanced” species tree estimates, such as tree  $T = ((a, b), (c, d))$  in Fig. 1, rather than unbalanced trees, such as tree  $S = (d, (b, (a, c)))$  in Fig. 1 (M. DeGiorgio, J. Syring, A.J. Eckert, A.I. Liston, R. Cronn, D.B. Neale, and N.A. Rosenberg, unpublished data).

In this paper, to further investigate the behavior of the MDC criterion, we study several of its mathematical properties. After introducing notation in Section 2, we describe in Section 3 the relationship between deep coalescence cost and the key concept of maximal subtrees. In Section 4, we determine a relationship between the deep coalescence cost and the rooted nearest-neighbor interchange (NNI) distance [10], [11], [12], which provides another measure of the amount of topological discordance between two trees. The problem of identifying gene trees that maximize the deep coalescence cost given a fixed species tree, and the dual problem of identifying species trees that maximize the deep coalescence cost given a gene tree, are investigated in Section 5.

## 2 NOTATION

We consider binary, rooted trees that are leaf-labeled and have at least two leaves. The set of all binary, rooted trees whose leaves are labeled by elements of a label set  $X$  is denoted by  $R(X)$ . For a tree  $T$ , let  $V(T)$  and  $\hat{V}(T)$  be the sets of nodes and internal (i.e., nonleaf) nodes of  $T$ , respectively. If an edge of  $T$  is incident to a leaf of  $T$ , it is called a pendant edge; otherwise, it is an internal edge. The sets of edges and internal edges of  $T$  are denoted, respectively, by  $E(T)$  and  $\hat{E}(T)$ . We denote by  $\rho(T)$  the root of  $T$ .

For a node  $v$  of tree  $T$ , let  $T(v)$  be the subtree of  $T$  induced by  $v$ , that is, the subtree rooted at  $v$  consisting of  $v$  and all proper descendants of  $v$  (node  $w$  is a proper descendant of  $v$  if  $w \neq v$  and  $v$  lies on the path from  $\rho(T)$  to  $w$ ). The cluster  $C_T(v)$  induced by  $v$  is defined as the set of leaves of  $T(v)$ , and we denote by  $n_v$  the number of elements of  $C_T(v)$ .

• C.V. Than is with the ZBIT, Department of Computer Science, University of Tuebingen, Sand 14, 72076 Tuebingen, Germany.

E-mail: cuong.van-than@uni-tuebingen.de, cvthan@stanford.edu.  
 • N.A. Rosenberg is with the Department of Biology, Stanford University, 371 Serra Mall, Stanford, CA 94305. E-mail: noahr@stanford.edu.

Manuscript received 19 Apr. 2012; revised 9 Oct. 2012; accepted 17 Oct. 2012; published online 23 Oct. 2012.

For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2012-04-0105. Digital Object Identifier no. 10.1109/TCBB.2012.133.

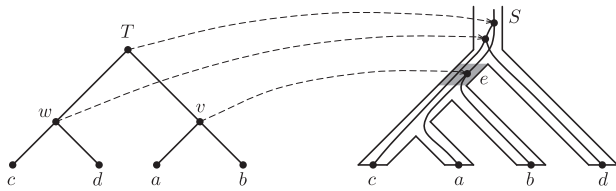


Fig. 1. Reconciling gene tree  $T$  within species tree  $S$ . In the figure, MRCA mappings between leaves of  $T$  and leaves of  $S$  are omitted, and for clearer illustration of the reconciliation process, the MRCAs of the internal nodes of  $T$  are placed along the edges of  $S$  and along the artificial edge above the root of  $S$  rather than at the internal nodes of  $S$ .

Edges of  $T$  are considered to be directed away from the root of  $T$ . Edge  $e$  of  $T$  with endpoints  $u$  and  $v$ , where  $u$  is the parent of  $v$ , is written as  $(u, v)$ . We call  $u$  and  $v$  the tail and head of  $e$ , respectively. Because a nonroot node  $v$  has only one parent  $u$ , it uniquely determines edge  $e = (u, v)$ . Therefore, for convenience we also refer to  $T(v)$ ,  $C_T(v)$ , and  $n_v$  as  $T(e)$ ,  $C_T(e)$ , and  $n_e$ , respectively.

Several properties of binary, rooted trees will be used throughout the paper. The number of trees in  $R(X)$  is  $1 \cdot 3 \cdots (2|X| - 3)$ , where  $|X|$  is the number of elements of  $X$ . A tree  $T \in R(X)$  has  $|X| - 1$  internal nodes and  $|X| - 2$  internal edges. Thus,  $|V(T)| = 2|X| - 1$  and  $|E(T)| = 2|X| - 2$ .

### 3 DEEP COALESCENCE COST

We assume incomplete lineage sorting [1], [8], [13] is the only process that can cause a gene tree to have a different topology from a species tree. Consider a specific species tree  $S$  and gene tree  $T$ , both binary and rooted, and with the same set of leaves. The deep coalescence cost for the pair consisting of  $T$  and  $S$  is computed as follows [1], [3]. Each node  $v$  of  $T$  is mapped to the most recent node of  $S$  (i.e., the farthest node from the root of  $S$ ) whose descendant leaf set contains the cluster  $C_T(v)$  induced by  $v$ . We denote by  $\text{MRCA}_S(v)$  the node to which  $v$  is mapped, and call it the most recent common ancestor (MRCA) of  $v$  in  $S$ . For an edge  $e$  of  $S$ , denote by  $c_e$  the number of internal nodes  $v$  of  $T$  that are mapped to nodes of the subtree  $S(e)$  induced by  $e$ . The number of *extra lineages* in  $e$  is defined as

$$\text{xl}(T, e) = n_e - c_e - 1. \quad (1)$$

The *deep coalescence cost* for the pair consisting of  $T$  and  $S$  is defined as the sum of  $\text{xl}(T, e)$  across all the edges of  $S$ , that is,

$$\text{dc}(T, S) = \sum_{e \in E(S)} \text{xl}(T, e). \quad (2)$$

Our definition of extra lineages arises from the way in which gene tree  $T$  is reconciled within the edges of  $S$  according to an MRCA mapping between  $T$  and  $S$ . For each leaf  $x$  of  $T$ , we place a gene lineage in the pendant edge of  $S$  incident to the leaf (of  $S$ ) labeled by  $x$ . We then recursively create a lineage for each internal node  $v$  of  $T$  by merging the two lineages for the two children of  $v$  in the edge of  $S$  with  $\text{MRCA}_S(v)$  as its head (or the edge above the root node  $\rho(S)$  if  $\text{MRCA}_S(v) = \rho(S)$ ). The reconciliation of  $T$  within  $S$  can also be viewed as the backward evolution of the lineages in the pendant edges of  $S$ , with the MRCA mapping

specifying where in  $S$  coalescences occur. For an edge  $e$  of  $S$ , if  $c_e$  internal nodes of  $T$  are mapped to nodes of subtree  $S(e)$ , then  $c_e$  coalescences occur along edge  $e$  and its proper descendant edges. Thus, the  $n_e$  lineages in the pendant edges of  $S(e)$  are merged into  $n_e - c_e$  lineages in edge  $e$ , and we say that edge  $e$  has  $n_e - c_e - 1$  extra lineages.

It can be verified from the reconciliation process above that the edge above the root of  $S$  has exactly one lineage. Thus, this edge has no extra lineages, and it need not be included in  $\text{dc}(T, S)$ . Notice also that  $c_e = 0$  for every pendant edge of  $S$ , because no internal nodes of  $T$  can be mapped to a leaf of  $S$ . The number of extra lineages in every pendant edge is  $1 - 0 - 1 = 0$ , and we can rewrite  $\text{dc}(T, S)$  as

$$\text{dc}(T, S) = \sum_{e \in \overset{\circ}{E}(S)} \text{xl}(T, e). \quad (3)$$

Fig. 1 illustrates the reconciliation of a gene tree  $T$  within a species tree  $S$ . Each of the two internal edges of  $S$  has one extra lineage, whereas there are no extra lineages in any of the pendant edges or in the edge above the root of  $S$ . Hence,  $\text{dc}(T, S) = 2$ .

Clearly,  $\text{dc}(T, S) = \text{dc}(S, T) = 0$  if and only if  $T$  and  $S$  have the same (labeled) topology. In general, however,  $\text{dc}(T, S) \neq \text{dc}(S, T)$ . For example, if  $T = (((a, b), c), d)$  and  $S = (((a, c), d), b)$ , then  $\text{dc}(T, S) = 3$  while  $\text{dc}(S, T) = 2$ . It is, therefore, important that in computing  $\text{dc}(T, S)$ , the gene tree  $T$  is reconciled within the species tree  $S$  and not vice versa.

We note that while the definition of the number of extra lineages by (1) arises naturally from the reconciliation of a gene tree  $T$  within a species tree  $S$  as described in [1], Zhang provided a different definition for this quantity [6], [7]. According to this definition, the nodes of  $T$  are mapped to the nodes of  $S$  by the MRCA mapping that we have described. An edge  $e$  of  $S$  has  $k_e - 1$  extra lineages if for exactly  $k_e$  edges  $(u, v)$  of  $T$ , edge  $e$  lies on the path from  $\text{MRCA}_S(u)$  to  $\text{MRCA}_S(v)$ . We will see in the next section that Zhang's definition and the definition in (1) are equivalent.

#### 3.1 Maximal Subtrees and the Number of Extra Lineages

Let  $T$  be a binary, rooted tree on  $X$  and let  $A$  be a nonempty subset of  $X$ . We say that a subtree  $T(v)$  induced by a node  $v$  of  $T$  is *A-maximal* if

1. the leaf set  $C_T(v)$  of  $T(v)$  is a subset of  $A$ , and
2. for any subtree  $t$  of  $T$  of which  $T(v)$  is a proper subtree, the leaf set of  $t$  is *not* a subset of  $A$ .

The concept of *A-maximality* is important throughout the article, and we describe it in detail. As an example, we compute subtrees of tree  $T$  in Fig. 1 that are *A-maximal*, where  $A = \{a, b, c\}$ . Cherry  $(a, b)$  is an *A-maximal* subtree of  $T$  because  $\{a, b\} \subseteq A$  and the leaf set of  $T$ —the only subtree of  $T$  of which  $(a, b)$  is a proper subtree—is not a subset of  $A$ . Similarly, leaf  $c$  is also an *A-maximal* subtree of  $T$ . None of the other subtrees of  $T$  is *A-maximal*.

We provide a simpler criterion for determining whether a subtree of  $T$  is *A-maximal*. Clearly,  $T$  is the only *X-maximal* subtree of  $T$ . Suppose that  $A$  is a proper, nonempty subset of  $X$ , and let  $T(v)$  be an *A-maximal*

subtree of  $T$ . Because  $C_T(v) \subseteq A$ ,  $C_T(v)$  is also a proper subset of  $X$  and hence,  $v$  is a nonroot node of  $T$ . Let  $u$  be the parent of  $v$ . As  $T(v)$  is a proper subtree of  $T(u)$ , the leaf set of  $T(u)$  (i.e.,  $C_T(u)$ ) is not a subset of  $A$  by condition 2. Conversely, suppose that  $C_T(v) \subseteq A$  and  $C_T(u) \not\subseteq A$ . Let  $t$  be a subtree of  $T$  that contains  $T(v)$  as a proper subtree. Then  $t$  is induced either by  $u$  or by a proper ancestor of  $u$ . It follows that the leaf set of  $t$  contains  $C_T(u)$ , implying that the leaf set of  $t$  is not a subset of  $A$ . By the two conditions above,  $T(v)$  is  $A$ -maximal. Therefore, we can say that  $T(v)$ , where  $v$  is a nonroot node of  $T$ , is  $A$ -maximal if and only if

1. the cluster  $C_T(v)$  induced by  $v$  is a subset of  $A$ , and
2. the cluster  $C_T(u)$  induced by the parent  $u$  of  $v$  is *not* a subset of  $A$ .

Consider a given species tree  $S$  and gene tree  $T$ . For an edge  $e$  of  $S$ , let  $k_e$  be the number of  $C_S(e)$ -maximal subtrees of  $T$ . Than and Nakhleh proved that [3]

$$\text{xl}(T, e) = k_e - 1. \quad (4)$$

Equation (4) agrees with Zhang's definition of extra lineages [6], [7]. If edge  $e$  of  $S$  lies on the path from  $\text{MRCA}_S(u)$  to  $\text{MRCA}_S(v)$ , where  $(u, v)$  is an edge of  $T$ , then  $\text{MRCA}_S(v)$  is a descendant of the head of  $e$  and  $\text{MRCA}_S(u)$  is an ancestor of the tail of  $e$ . This means that  $T(v)$  is a maximal subtree of  $T$  with respect to  $C_S(e)$ . Conversely, if  $T(v)$  is a  $C_S(e)$ -maximal subtree of  $T$ , then by definition,  $C_T(v) \subseteq C_S(e)$  and  $C_T(u) \not\subseteq C_S(e)$ . Hence,  $\text{MRCA}_S(v)$  is a descendant of the head of  $e$ , while  $\text{MRCA}_S(u)$  is an ancestor of the tail of  $e$ . Thus, each  $C_S(e)$ -maximal subtree of  $T$  corresponds to exactly one edge  $(u, v)$  of  $T$  for which edge  $e$  lies on the path from  $\text{MRCA}_S(u)$  to  $\text{MRCA}_S(v)$ . That is, the number of extra lineages in edge  $e$  by Zhang's definition is also  $k_e - 1$ .

As an illustration of (4), consider the shaded edge  $e$  of the species tree  $S$  in Fig. 1. Edge  $e$  induces cluster  $C_S(e) = \{a, b, c\}$ , and as noted earlier, only leaf  $c$  and cherry  $(a, b)$  of  $T$  are  $C_S(e)$ -maximal. Equation (4) gives  $\text{xl}(T, e) = 2 - 1 = 1$ , which agrees with the illustration in Fig. 1 that edge  $e$  has one extra lineage.

Equation (4) eliminates the need for the MRCA mapping between the nodes of  $T$  and the nodes of  $S$  in computing  $\text{dc}(T, S)$ . It also shows that the number of extra lineages in species tree edge  $e$  depends only on the cluster  $C_S(e)$  (and the gene tree), and not on the labeled topology of the subtree  $S(e)$ . In other words, we can associate with each  $A \subseteq X$  the cost  $\text{xl}(T, A)$ . This result is the basis for a dynamic programming algorithm for identifying a species tree with the minimum deep coalescence cost for a collection of gene trees [3].

#### 4 DEEP COALESCENCE COST AND ROOTED-NNI DISTANCE

In this section, we look at some relationships between the deep coalescence cost and the *rooted-NNI distance*. A single rooted-NNI move, when applied to a subtree  $t = ((A, B), C)$  of a tree  $T$ , transforms  $T$  into a new tree with the subtree  $t$  substituted by either  $t' = (A, (B, C))$  or  $t'' = (B, (A, C))$  (Fig. 2); here, letters  $A$ ,  $B$ , and  $C$  denote subtrees of  $T$ . For two trees  $T$  and  $S$  in  $R(X)$ , the rooted-NNI

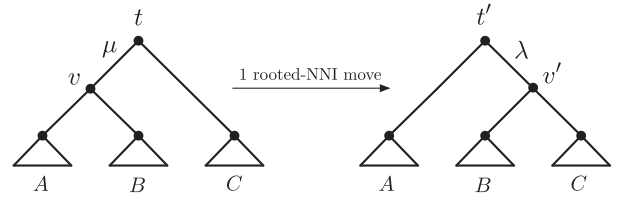


Fig. 2. A rooted nearest-neighbor interchange move. The tree  $t' = (A, (B, C))$  (right) can be obtained from the tree  $t = ((A, B), C)$  (left) by one rooted-NNI move. The subtree  $B$  is pruned from  $t$  then reattached to  $t$  as a sibling of  $C$ . An alternative rooted-NNI move, in which  $A$  is cut instead of  $B$  and rejoined to  $t$  as a sibling of  $C$ , creates another tree  $t'' = (B, (A, C))$ .

distance between  $T$  and  $S$ , denoted henceforth  $d_{\text{nni}}(T, S)$ , is the minimum number of rooted-NNI moves required to convert  $T$  into  $S$ . The rooted-NNI distance is a distance metric on the set  $R(X)$  [12] (see also [10]).

**Theorem 1.** *Let  $T$  and  $S$  be two binary, rooted trees on  $X$ . Then,  $\text{dc}(T, S) = 1$  if and only if  $d_{\text{nni}}(T, S) = 1$ .*

**Proof.** Suppose that  $d_{\text{nni}}(T, S) = 1$ . Then  $T$  and  $S$  must have at least three leaves. We can assume that  $T$  has a subtree  $t = ((A, B), C)$ , where  $A$ ,  $B$ , and  $C$  represent subtrees, and that  $S$  is obtained from  $T$  by substituting  $t$  with  $t' = (A, (B, C))$ . Let  $\lambda$  be the edge of  $S$  that induces subtree  $(B, C)$  (Fig. 2). It is easy to see that  $B$  and  $C$  are the only subtrees of gene tree  $T$  that are  $C_S(\lambda)$ -maximal. By (4),  $\text{xl}(T, \lambda) = 1$ . Further, it can be checked that for any edge  $e$  of  $S$  other than  $\lambda$ , there is a unique edge in  $T$  that induces the same cluster as  $C_S(e)$ . It follows that there is only one  $C_S(e)$ -maximal subtree of  $T$ . By (4) again,  $\text{xl}(T, e) = 0$ . Thus, we have  $\text{dc}(T, S) = \text{xl}(T, \lambda) + \sum_{e \neq \lambda} \text{xl}(T, e) = 1$ .

For the converse, suppose that  $\text{dc}(T, S) = 1$ . Trees  $T$  and  $S$  are not identical, for otherwise  $\text{dc}(T, S) = 0$ . Let  $\lambda$  be an (internal) edge of  $S$  such that  $C_S(\lambda)$  is not an induced cluster of  $T$ . Then there must be at least two subtrees of  $T$  that are  $C_S(\lambda)$ -maximal, and hence  $\text{xl}(T, \lambda) \geq 1$  by (4). Because  $\text{dc}(T, S) = 1$ ,  $C_S(\lambda)$  is the only cluster of  $S$  that is not a cluster of  $T$  (and  $\text{xl}(T, \lambda) = 1$ ). This in turn implies that  $T$  also has only one cluster, say,  $C_T(\mu)$  induced by edge  $\mu$ , that is not a cluster of  $S$  (note that  $S$  and  $T$  induce the same number of clusters, as they have the same number of edges). Contracting either edge  $\lambda$  in  $S$  or edge  $\mu$  in  $T$  results in the same tree, which we denote by  $R$ . Let  $w$  be the only ternary node of  $R$ , and let  $A$ ,  $B$ , and  $C$  be the three subtrees of  $R$  attached to  $w$ . To obtain  $S$  and  $T$ , these subtrees must be regrouped in two different ways. Without loss of generality, we can assume that  $A$  and  $B$  are grouped as siblings in  $T$ , whereas  $B$  and  $C$  are grouped as siblings in  $S$  (Fig. 2). Clearly, one rooted-NNI move is required to transform  $T$  to  $S$  and vice versa, and  $d_{\text{nni}}(T, S) = 1$ .  $\square$

For each internal edge of a tree  $T \in R(X)$ , we can obtain from  $T$  two different trees by one rooted-NNI move. The number of internal edges of  $T$  is  $|X| - 2$ , and hence, there are  $2(|X| - 2)$  trees  $S$  with  $d_{\text{nni}}(T, S) = 1$  [10], [11]. Note also that the rooted-NNI distance is symmetric, as it is a metric on the set  $R(X)$ . We have the following corollaries.

**Corollary.** Let  $T$  and  $S$  be two binary, rooted tree on  $X$ . Then  $\text{dc}(T, S) = 1$  if and only if  $\text{dc}(S, T) = 1$ .

**Corollary.** Let  $T$  be a binary, rooted tree on  $X$ . Then the number of trees  $S \in R(X)$  with  $\text{dc}(T, S) = \text{dc}(S, T) = 1$  is  $2(|X| - 2)$ .

By Theorem 1, if  $d_{\text{nni}}(T, S) = 2$ , then  $\text{dc}(T, S) \geq 2$ , and if  $\text{dc}(T, S) = 2$ , then  $d_{\text{nni}}(T, S) \geq 2$ . It is, therefore, tempting to conjecture that  $\text{dc}(T, S) \leq d_{\text{nni}}(T, S)$  or  $d_{\text{nni}}(T, S) \leq \text{dc}(T, S)$ , but neither inequality holds in general. Consider two trees,  $T = ((a, b), (c, d))$  and  $S = ((a, c), (b, d))$ . We have  $\text{dc}(T, S) = 2$ , while at least three rooted-NNI moves are needed to convert  $T$  into  $S$ . On the other hand, for the pair of trees  $T' = (c, (a, (b, d)))$  and  $S' = (d, (b, (c, a)))$ , we have  $\text{dc}(T', S') = 3$ , while  $d_{\text{nni}}(T', S') = 2$ . However, the next result demonstrates that a weaker relationship between the deep coalescence cost and the rooted-NNI distance does exist.

**Theorem 2.** Let  $S$  and  $S'$  be two binary, rooted species on  $X$ . Assume that  $d_{\text{nni}}(S, S') = 1$ . Then for any gene tree  $T \in R(X)$ ,

$$|\text{dc}(T, S) - \text{dc}(T, S')| \leq |X| - 2. \quad (5)$$

**Proof.** Because  $d_{\text{nni}}(S, S') = 1$ , we can assume that  $S'$  is obtained from  $S$  by substituting a subtree  $((A, B), C)$  of  $S$  with  $(A, (B, C))$ , where  $A, B$ , and  $C$  represent subtrees of  $S$  (see Fig. 2, considering  $t$  and  $t'$  as subtrees of  $S$  and  $S'$ , respectively). Hence, every cluster induced by  $S$  appears in  $S'$ , except for the cluster induced by edge  $\mu$  in  $S$ . Conversely, every cluster induced by  $S'$  also appears in  $S$ , except for the cluster induced by edge  $\lambda$  in  $S'$ . Applying (1) and (2), we have

$$\begin{aligned} |\text{dc}(T, S) - \text{dc}(T, S')| &= |\text{xl}(T, \mu) - \text{xl}(T, \lambda)| \\ &\leq \max\{\text{xl}(T, \mu), \text{xl}(T, \lambda)\}, \end{aligned}$$

where the last inequality follows from the fact that both  $\text{xl}(T, \mu)$  and  $\text{xl}(T, \lambda)$  are nonnegative. The number of lineages in edge  $\mu$  is at most the total number of leaves in  $A$  and  $B$ , which is at most  $|X| - 1$  because subtree  $C$  has at least one leaf. Similarly, the number of lineages in edge  $\lambda$  is at most  $|X| - 1$ . Therefore,  $|\text{dc}(T, S) - \text{dc}(T, S')| \leq |X| - 2$ .  $\square$

**Theorem 3.** Let  $T$  and  $T'$  be two binary, rooted gene trees on  $X$ . Assume that  $d_{\text{nni}}(T, T') = 1$ . For any species tree  $S \in R(X)$ ,

$$|\text{dc}(T, S) - \text{dc}(T', S)| \leq |X| - 2. \quad (6)$$

**Proof.** We have

$$\begin{aligned} |\text{dc}(T, S) - \text{dc}(T', S)| &= \left| \sum_{e \in \hat{E}(S)} \text{xl}(T, e) - \sum_{e \in \hat{E}(S)} \text{xl}(T', e) \right| \\ &\leq \sum_{e \in \hat{E}(S)} |\text{xl}(T, e) - \text{xl}(T', e)| \\ &= \sum_{e \in \hat{E}(S)} |c_e - c'_e|, \quad (\text{by (1)}) \end{aligned}$$

where  $c_e$  and  $c'_e$ , respectively, denote the numbers of internal nodes of  $T$  and  $T'$  whose MRCA's in  $S$  are nodes of the subtree  $S(e)$  induced by  $e$ . Because  $|\hat{E}(S)| = |X| - 2$ , it is sufficient to show that for each internal edge  $e$  of  $S$ ,

$$|c_e - c'_e| \leq 1.$$

Because  $d_{\text{nni}}(T, T') = 1$ , we can assume that  $T'$  is obtained from  $T$  by substituting a subtree  $((A, B), C)$  of  $T$  with  $(A, (B, C))$ , where  $A, B$ , and  $C$  represent subtrees (Fig. 2). Let  $v$  be the node of  $T$  that induces the subtree  $(A, B)$ , and let  $v'$  be the node of  $T'$  that induces  $(B, C)$ . Except for clusters  $C_T(v)$  of  $T$  and  $C_{T'}(v')$  of  $T'$ , every cluster of  $T$  is a cluster of  $T'$  and vice versa. Thus, for every node  $u$  of  $T$  other than  $v$ , there exists a unique node  $u' \neq v'$  in  $T'$  such that both  $u$  and  $u'$  induce the same cluster. This implies that  $\text{MRCA}_S(u) = \text{MRCA}_S(u')$  if  $u \neq v$  and  $u' \neq v'$ , and only  $\text{MRCA}_S(v)$  and  $\text{MRCA}_S(v')$  can differ. As a consequence, the value of  $c_e - c'_e$  depends only on whether  $\text{MRCA}_S(v)$  and  $\text{MRCA}_S(v')$  are nodes of  $S(e)$ . Therefore,

1. If both  $\text{MRCA}_S(v)$  and  $\text{MRCA}_S(v')$  are nodes of  $S(e)$ , or if neither is a node of  $S(e)$ , then  $c_e = c'_e$ .
2. If  $\text{MRCA}_S(v)$  is a node of  $S(e)$ , but  $\text{MRCA}_S(v')$  is not, then  $c_e = c'_e + 1$ .
3. If  $\text{MRCA}_S(v')$  is a node of  $S(e)$ , but  $\text{MRCA}_S(v)$  is not, then  $c'_e = c_e + 1$ .

The desired claim holds in all three cases.  $\square$

It can be seen that the upper bound in Theorem 2 is tight. Consider an  $n$ -leaf caterpillar species tree  $S = (\dots ((1, 2), 3), \dots, n)$ , and let  $S'$  be a species tree obtained from  $S$  by applying one rooted-NNI move that makes leaves  $n - 1$  and  $n$  siblings (i.e.,  $S' = (\dots ((1, 2), 3), \dots, n - 2), (n - 1, n))$ ). Let  $T_1 = (\dots (((n - 1, n), 1), 2), \dots, n - 2)$ . By direct calculation, we have  $\text{dc}(T_1, S) = (n - 2)(n - 1)/2$  and  $\text{dc}(T_1, S') = (n - 3)(n - 2)/2$ . Hence,  $\text{dc}(T_1, S) - \text{dc}(T_1, S') = n - 2$ .

As for Theorem 3, consider caterpillar gene trees  $T_2 = (\dots ((1, n), 2), \dots, n - 1)$  and  $T'_2 = (\dots (((1, 2), n), 3), \dots, n - 1)$  that are one rooted-NNI move apart. Again by direct computation, we have  $\text{dc}(T_2, S) = (n - 2)(n - 1)/2$  and  $\text{dc}(T'_2, S) = (n - 3)(n - 2)/2$ . Hence,  $\text{dc}(T_2, S) - \text{dc}(T'_2, S) = n - 2$ , and the bound in Theorem 3 is also tight.

## 5 TREES WITH MAXIMUM DEEP COALESCENCE COST

In this section, we solve the following problem: given a fixed species tree  $S$  (or a fixed gene tree  $T$ ), which gene trees (or species trees) have the maximum deep coalescence cost?

### 5.1 Fixed Species Trees

We first consider the case in which we are given a fixed species tree  $S \in R(X)$ ,  $|X| = n \geq 2$ . We derive an upper bound for  $\text{dc}(T, S)$  over all gene trees  $T \in R(X)$ . We also derive a formula for the number of gene trees  $T$  with  $\text{dc}(T, S)$  equal to that upper bound.

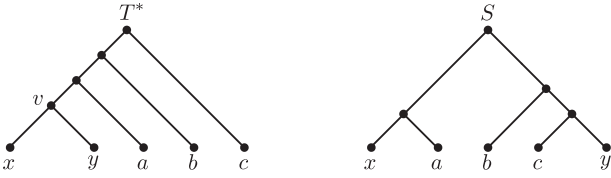


Fig. 3. An example of the maximum deep coalescence cost for a fixed species tree. Given the species tree  $S$  (right), the deep coalescence cost for reconciling the gene tree  $T^*$  (left) achieves the maximum deep coalescence cost  $\text{dc}(T^*, S) = 4$  over all gene trees.

### 5.1.1 Upper Bound of $\text{dc}(T, S)$

From (1) and (2),  $\text{dc}(T, S)$  is bounded above by  $m_s(S) = \sum_{e \in E(S)} (n_e - 1) = -(2n - 2) + \sum_{e \in E(S)} n_e$ . We now construct a tree  $T^*$  with  $\text{dc}(T^*, S) = m_s(S)$ . Let  $x$  be a leaf in one subtree of the root of  $S$ , and let  $y$  be a leaf in the other subtree of the root. Let  $T^*$  be a caterpillar tree on  $X$  whose only cherry is  $(x, y)$ , and let  $v$  be the node of  $T^*$  that induces this cherry (Fig. 3). Because leaves  $x$  and  $y$  of the species tree  $S$  appear in different subtrees of the root of  $S$ ,  $\text{MRCA}_S(v)$  is the root of  $S$ . It follows that the MRCA of any internal node of  $T^*$  in  $S$  is also the root of  $S$ . Consequently, for each edge  $e$  of  $S$ ,  $c_e = 0$ . By (1),  $\text{xl}(T^*, e) = n_e - 1$ , and by (2),  $\text{dc}(T^*, S) = m_s(S)$ . We have proven

**Lemma 4.** For a given species tree  $S \in R(X)$ , the deep coalescence cost for reconciling a gene tree  $T \in R(X)$  is bounded above by

$$m_s(S) = -(2n - 2) + \sum_{e \in E(S)} n_e. \quad (7)$$

The depth of a node  $v$  in a tree  $S$ , denoted henceforth by  $\ell(v)$ , is defined as the number of edges in the (unique) path from the root of  $S$  to  $v$ . The external path length of  $S$ ,  $\text{epl}(S)$ , is defined as the sum  $\sum_{x \in X} \ell(x)$ .

**Lemma 5.** Let  $S$  be a tree in  $R(X)$ , where  $|X| = n$ . Then

$$\sum_{e \in E(S)} n_e = \text{epl}(S) \leq \frac{n(n+1)}{2} - 1, \quad (8)$$

with equality if and only if  $S$  is a caterpillar tree.

**Proof.** The first part of the lemma follows by noting that each leaf  $x$  appears in exactly  $\ell(x)$  proper subtrees of  $S$ , and hence it contributes exactly  $\ell(x)$  to the sum  $\sum_{e \in E(S)} n_e$  (e.g., [14], [15]). The second part was proven by Klein and Wood [16] (see also [17, Section 2.3.4.5]).  $\square$

The following theorem is a direct consequence of Lemmas 4 and 5.

**Theorem 6.** For any pair consisting of a species tree  $S$  and a gene tree  $T$  in  $R(X)$ ,  $|X| = n$ ,

$$\text{dc}(T, S) \leq m_s(S) \leq \frac{(n-2)(n-1)}{2}. \quad (9)$$

### 5.1.2 Number of Gene Trees with Maximum Deep Coalescence Cost

When  $S$  is not a caterpillar tree, gene trees  $T$  need not be caterpillar trees to have  $\text{dc}(T, S) = m_s(S)$  in (7). For

example, the gene tree obtained from  $T^*$  in Fig. 3 by pruning leaf  $a$  and reattaching it to  $T^*$  as a sibling of leaf  $c$  also has deep coalescence cost  $m_s(S) = \text{dc}(T^*, S) = 4$ . We establish conditions for gene trees  $T$  to have  $\text{dc}(T, S) = m_s(S)$  and derive a formula for the number of these gene trees.

Denote by  $S_\ell$  and  $S_r$  the two subtrees of the root of  $S$ . A cherry of  $T$  is called a *left-right cherry* with respect to  $S$  if it has one leaf each from  $S_\ell$  and  $S_r$ .

**Lemma 7.** Let  $S \in R(X)$  be a given species tree. Then a gene tree  $T \in R(X)$  has  $\text{dc}(T, S) = m_s(S)$  if and only if every cherry of  $T$  is a left-right cherry with respect to  $S$ .

**Proof.** For the necessary condition, suppose otherwise that  $T$  has a cherry whose leaves are both from  $S_\ell$  (the case in which they are from  $S_r$  is similar). Let  $v$  be the node that induces this cherry. Then  $\text{MRCA}_S(v)$  is some node of  $S_\ell$ , and hence is a proper descendant of the root of  $S$ . Let  $\lambda$  be the edge of  $S$  whose head is  $\text{MRCA}_S(v)$ . By the definition of  $c_\lambda$ , we have  $c_\lambda \geq 1$ . From (1),  $\text{xl}(T, \lambda) = n_\lambda - c_\lambda - 1 < n_\lambda - 1$ , and it follows that  $\text{dc}(T, S) = \sum_{e \in E(S)} \text{xl}(T, e) < m_s(S)$ .

For the converse, if every cherry of  $T$  is left-right with respect to  $S$ , then  $\text{MRCA}_S(v)$  is the root of  $S$  for every internal node  $v$  of  $T$ . Thus,  $c_e = 0$  for each edge of  $S$ , and by (2),  $\text{dc}(T, S) = m_s(S)$ .  $\square$

**Lemma 8.** Let  $k$  and  $n - k$  be the numbers of leaves in  $S_\ell$  and  $S_r$ , respectively. Then a gene tree  $T \in R(X)$  that has  $\text{dc}(T, S) = m_s(S)$  cannot have more than  $\min\{k, n - k\}$  cherries.

**Proof.** By Lemma 7, each cherry of  $T$  is a left-right cherry with respect to  $S$ . Because  $S_\ell$  has  $k$  leaves and  $S_r$  has  $n - k$  leaves, at most  $\min\{k, n - k\}$  left-right cherries can be formed.  $\square$

Lemmas 7 and 8 reduce the problem of counting the number of gene trees  $T$  with  $\text{dc}(T, S) = m_s(S)$  to the problem of counting for each  $i = 1, \dots, \min\{k, n - k\}$  the number of gene trees that have  $i$  cherries, each of which is left-right with respect to  $S$ . We count these gene trees by using a bijection of [18] between binary, rooted trees and perfect matchings.

A perfect matching on  $2n - 2$  points is simply a set of  $n - 1$  unordered pairs of these points. For ease of description of the bijection, assume that the leaves of a tree  $T$  are labeled by integers  $1, 2, \dots, n$ . The internal nodes of  $T$ , excluding the root, are assigned integers  $n + 1, \dots, 2n - 2$  by repeating the following procedure.

1. Consider the set  $U$  of unlabeled internal nodes both of whose children have already been labeled.
2. Find among the children of the elements of  $U$  node  $w$  that has the smallest label. Let  $v$  be the parent of  $w$ .
3. Label  $v$  with the next available integer.

For each internal node of  $T$ , including the root, we form an unordered pair of the labels of its two children. The set of all these pairs is the perfect matching for  $T$  (Fig. 4).

Diaconis and Holmes proved that each binary, rooted tree with leaves labeled by integers  $1, 2, \dots, n$  corresponds to exactly one perfect matching on points  $1, 2, \dots, 2n - 2$  according to the procedure above [18]. In fact, the number of perfect matchings on  $2n - 2$  points is



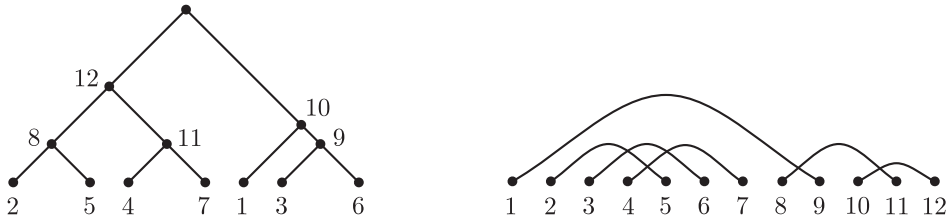


Fig. 4. A correspondence between a binary, rooted tree, and a perfect matching. The tree has leaves labeled by  $1, 2, \dots, 7$  (left), and the perfect matching (right) is on points  $1, 2, \dots, 12$ . The internal nodes of the tree are labeled according to the procedure described in the text. The matching is formed by pairing the two children of each internal node of the tree.

$$\frac{1}{(n-1)!} \binom{2n-2}{2} \binom{2n-4}{2} \cdots \binom{2}{2} = (2n-3)!!,$$

which is also the number of binary, rooted trees with  $n$  leaves.

**Lemma 9.** Let the numbers of leaves in  $S_\ell$  and  $S_r$  be  $k$  and  $n-k$ , respectively. The number of trees  $T$  that have  $i$  cherries, each of which is left-right with respect to  $S$ , is

$$\frac{(n-2)!i}{2^{i-1}} \binom{k}{i} \binom{n-k}{i}. \quad (10)$$

**Proof.** We use the tree-matching bijection to construct trees  $T$  with  $i$  left-right cherries with respect to  $S$ . Noting that in the bijection, points  $1, 2, \dots, n$  are assigned to the leaves of  $T$ , we divide the set of points  $1, 2, \dots, 2n-2$  into three subsets.

1.  $P_\ell$  is the set of  $k$  points corresponding to the leaves in  $S_\ell$ .
2.  $P_r$  is the set of  $n-k$  points corresponding to the leaves in  $S_r$ .
3.  $Q$  is the set of the  $n-2$  remaining points:  $n+1, \dots, 2n-2$ .

A tree  $T$  is built in three steps. First, we create  $i$  left-right cherries by pairing  $i$  points from  $P_\ell$  with  $i$  points from  $P_r$ . There are  $\binom{k}{i}$   $i$ -subsets of  $P_\ell$  and  $\binom{n-k}{i}$   $i$ -subsets of  $P_r$ . For a fixed  $i$ -subset of  $P_\ell$  and a fixed  $i$ -subset of  $P_r$ , there are  $i!$  possible pairings: the first element from the subset of  $P_\ell$  pairs with one of  $i$  choices from the subset of  $P_r$ , the second element pairs with one of  $i-1$  choices, and so on. Thus, the number of possible sets of  $i$  left-right cherries is

$$i! \binom{k}{i} \binom{n-k}{i}.$$

After this pairing,  $k-i$  and  $n-k-i$  points remain in  $P_\ell$  and  $P_r$ , respectively. These  $(k-i) + (n-k-i) = n-2i$  points must be paired with  $n-2i$  points in  $Q$ , because we require that  $T$  has exactly  $i$  cherries. Using the same argument as in the preceding paragraph, the number of pairings in this step is

$$(n-2i)! \binom{n-2i}{n-2i} \binom{n-2}{n-2i} = \frac{(n-2)!}{(2i-2)!}.$$

In the third and final step,  $(n-2) - (n-2i) = 2i-2$  points in  $Q$  are paired. Any pairing of these points is allowed, as these points correspond to internal nodes of tree  $T$ . The number of different pairings is  $(2i-3)!!$ , the

number of binary, rooted trees with  $i$  leaves (by the tree-matching bijection).

The total number of trees that have  $i$  left-right cherries with respect to  $S$  is

$$\begin{aligned} & i! \binom{k}{i} \binom{n-k}{i} \cdot \frac{(n-2)!}{(2i-2)!} \cdot (2i-3)!! \\ &= i! \binom{k}{i} \binom{n-k}{i} \cdot \frac{(n-2)!}{(2i-2)!} \cdot \frac{(2i-2)!}{2^{i-1}(i-1)!} \\ &= \frac{(n-2)!i}{2^{i-1}} \binom{k}{i} \binom{n-k}{i}. \end{aligned}$$

□

**Theorem 10.** Let the numbers of leaves in  $S_\ell$  and  $S_r$  be  $k$  and  $n-k$ , respectively. For a fixed species tree  $S \in R(X)$ , the number of gene trees  $T \in R(X)$  with  $\text{dc}(T, S) = m_s(S)$  is

$$\varphi_s(S) = 2(n-2)! \sum_{i=1}^{\min\{k, n-k\}} i 2^{-i} \binom{k}{i} \binom{n-k}{i}. \quad (11)$$






















































































**Proof.** The theorem follows immediately from Lemmas 7, 8, and 9. □

To illustrate the theorem, we count the number of trees  $T$  with  $\text{dc}(T, S) = m_s(S) = 4$  for the species tree  $S$  in Fig. 3. Because  $T$  has five leaves, we need to consider only two cases:  $T$  has one cherry (i.e., it is a caterpillar tree) and  $T$  has two cherries. In the first case, one leaf of the cherry is from  $\{a, x\}$ , and the other is from  $\{b, c, y\}$ . Hence, six different cherries can be formed. For a fixed one of these cherries, say,  $(x, y)$ , the remaining leaves of a caterpillar tree  $T$  can be assigned labels  $a, b$ , and  $c$  in  $3! = 6$  ways. The number of trees in the first case is therefore  $6 \cdot 6 = 36$ .

For the second case,  $T$  has two cherries, one containing  $a$  and the other containing  $x$  (by Lemma 7). Six possible sets of two such cherries exist. For a fixed set of two cherries, say,  $\{(a, b), (x, y)\}$ , we can arrange the two cherries and leaf  $c$  in three different ways to construct  $T$ . Thus, the number of trees in this case is  $6 \cdot 3 = 18$ . The total number of trees with  $\text{dc}(T, S) = 4$  is  $36 + 18 = 54$ . Simple calculation verifies that (11) also gives this number.

Let  $S$  be a fixed species tree on  $X$ , and let  $S'$  be another species tree that has the same shape (i.e., unlabeled topology) as  $S$ . There exists a permutation  $\pi$  on  $X$  such that  $S'$  is obtained from  $S$  by relabeling each leaf  $x$  of  $S$  with  $\pi(x)$ . Suppose that  $T$  is a gene tree that has  $\text{dc}(T, S) = m_s(S)$ . Let  $T'$  be a tree obtained from  $T$  by relabeling the leaves of  $T$

TABLE 1  
The Maximum Deep Coalescence Cost,  $m_s(S)$ , and the Number of Gene Trees with Maximum Cost,  $\varphi_s(S)$ , for an Arbitrarily Chosen Labeling of Fixed Species Tree Shapes with  $1 \leq n \leq 9$  Leaves

$n \leq 5$	 0, 1	 0, 1	 1, 2	 3, 6	 2, 10	 6, 24	 5, 24	 4, 54				
$n = 6$	 10, 120	 9, 120	 8, 120	 7, 336	 6, 336	 6, 450						
$n = 7$	 15, 720	 14, 720	 13, 720	 12, 720	 11, 720	 11, 720	 11, 2400	 10, 2400	 9, 2400	 9, 3960	 8, 3960	
$n = 8$	 21, 5040	 20, 5040	 19, 5040	 18, 5040	 17, 5040	 17, 5040	 17, 5040	 16, 5040	 15, 5040	 15, 5040	 14, 5040	 16, 19440
$n = 9$	 28, 40320	 27, 40320	 26, 40320	 25, 40320	 24, 40320	 24, 40320	 24, 40320	 23, 40320	 22, 40320	 22, 40320	 21, 40320	 23, 40320
	 22, 40320	 21, 40320	 20, 40320	 19, 40320	 19, 40320	 20, 40320	 19, 40320	 18, 40320	 19, 40320	 18, 40320	 17, 40320	 22, 176400
	 21, 176400	 20, 176400	 19, 176400	 18, 176400	 18, 176400	 18, 176400	 17, 176400	 16, 176400	 16, 176400	 15, 176400	 18, 393120	 17, 393120
	 16, 393120	 15, 393120	 14, 393120	 14, 393120	 16, 567000	 15, 567000	 14, 567000	 15, 567000	 14, 567000	 15, 567000	 14, 567000	 13, 567000

For each species tree shape  $S$ , quantities appear as an ordered pair  $(m_s(S), \varphi_s(S))$ . Species tree shapes are ordered according to their Furnas ranks.

according to permutation  $\pi$ . It can be seen that  $dc(T', S') = dc(T, S)$  [9]. Thus, the maximum cost  $m_s(S)$  and function  $\varphi_s(S)$  depend only on the shape of  $S$ , not on the leaf labels of  $S$ . This observation can be verified from (7) and (11).

### 5.1.3 Properties of $m_s(S)$ and $\varphi_s(S)$ for Small Trees

In this section, we examine the properties of  $m_s(S)$  and  $\varphi_s(S)$  in relation to the Furnas rank of  $S$  [20], denoted henceforth as  $\text{rank}_F(S)$ . Tree shapes with the same number of leaves are assigned consecutive positive integers, starting from 1. The procedure for assigning ranks to tree shapes is

recursive. There is only one tree shape with one leaf, and it has rank 1. For a tree shape  $S$  with at least two leaves, we designate  $S_\ell$  and  $S_r$  as the left and right subtrees of the root of  $S$  such that  $S_\ell$  has fewer leaves than  $S_r$ , or  $\text{rank}_F(S_\ell) \leq \text{rank}_F(S_r)$  in the case that  $S_\ell$  and  $S_r$  have the same number of leaves. For two tree shapes  $S$  and  $S'$ , if  $S_\ell$  has fewer leaves than  $S'_\ell$ , then  $\text{rank}_F(S) < \text{rank}_F(S')$ . If  $S_\ell$  and  $S'_\ell$  have the same number of leaves and  $\text{rank}_F(S_\ell) < \text{rank}_F(S'_\ell)$ , then  $\text{rank}_F(S) < \text{rank}_F(S')$ . Otherwise,  $S_\ell$  and  $S'_\ell$  have the same shape, and  $\text{rank}_F(S) < \text{rank}_F(S')$  if  $\text{rank}_F(S_r) < \text{rank}_F(S'_r)$ .

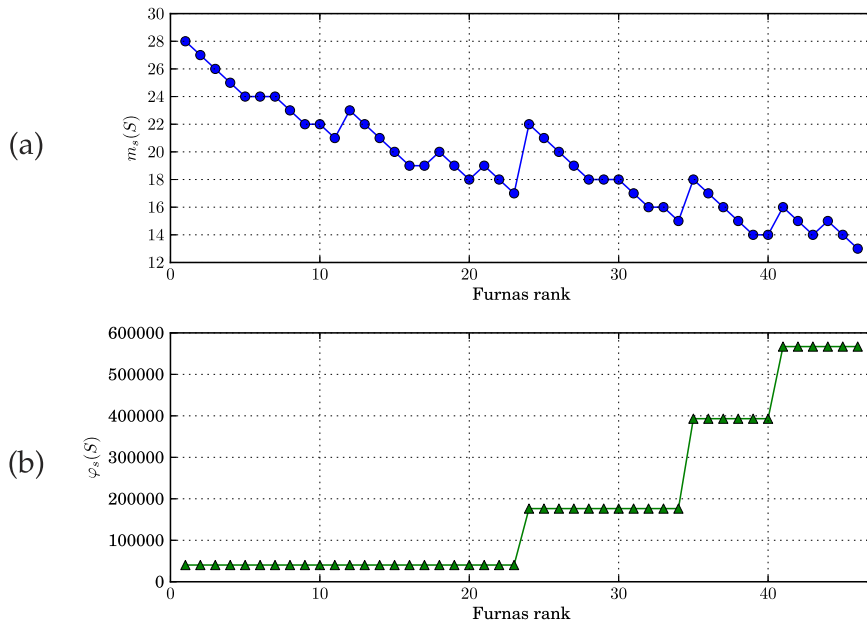


Fig. 5. Plots of  $m_s(S)$ , the maximum deep coalescence cost for fixed  $S$ , and  $\varphi_s(S)$ , the number of gene trees  $T$  with  $dc(T, S) = m_s(S)$ , for all 46 species tree shapes with nine leaves. The species tree shapes are ordered according to their Furnas ranks.

Caterpillar trees always have rank 1, and “most-balanced” trees—those in which for every internal node, the numbers of leaves in its left and right subtrees differ by at most one—have the largest ranks. In general, trees with higher  $\text{rank}_F$  are more balanced than trees with smaller  $\text{rank}_F$  (e.g., [20]).

The maximum cost  $m_s(S)$  and the number  $\varphi_s(S)$  of gene trees that achieve this maximum appear in Table 1 for every species tree shape with nine or fewer leaves. Examining the relationship between  $m_s(S)$  and  $\text{rank}_F$  for trees in the table, we see that trees with higher  $\text{rank}_F$  generally have lower  $m_s(S)$  (Fig. 5a). However, the relationship between  $m_s(S)$  and  $\text{rank}_F(S)$  is not completely monotonic, as shown in Fig. 5 and Table 1. Consider tree shapes with nine leaves, for example. Tree shape 23 has  $m_s(S) = 17$ , and tree shape 24 has  $m_s(S) = 22$ . Note that the right and left subtrees of shape 23 are a perfect balanced eight-leaf tree and a single-leaf tree, whereas the corresponding subtrees of shape 24 are a seven-leaf caterpillar and a cherry. From (7), we can see that the caterpillar subtree of shape 24 accounts for the increase in the maximum deep coalescence cost over that of shape 23.

As for  $\varphi_s(S)$ , Fig. 5b suggests that for nine taxa, it is monotonically nondecreasing with  $\text{rank}_F(S)$ . It can be proven that this monotonicity holds in general.

**Proposition 11.** *For two species tree shapes  $S$  and  $S'$  with the same number of leaves,  $\varphi_s(S) \leq \varphi_s(S')$  if  $\text{rank}_F(S) < \text{rank}_F(S')$ .*

**Proof.** By our designation, the left subtree  $S_\ell$  of each tree shape  $S$  has at most  $\lfloor n/2 \rfloor$  leaves. Further,  $\text{rank}_F(S) < \text{rank}_F(S')$  implies that the number of leaves in  $S_\ell$  is at most the number of leaves in  $S'_\ell$ . From (11),  $\varphi_s(S) = \varphi_s(S')$  if  $S_\ell$  and  $S'_\ell$  have the same number of leaves. Hence, it only remains to show that if  $S_\ell$  has fewer leaves than  $S'_\ell$ , then  $\varphi_s(S) < \varphi_s(S')$ . We have

$$\begin{aligned} \frac{\binom{k}{i} \binom{n-k}{i}}{\binom{k-1}{i} \binom{n-k+1}{i}} &= \frac{k}{k-i} \frac{n-k-i+1}{n-k+1} \\ &= \frac{k(n-k+1) - ki}{k(n-k+1) - (n-k+1)i}, \end{aligned}$$

which is greater than 1 if  $k \leq \lfloor n/2 \rfloor$ . Consequently, if  $2 \leq k \leq \lfloor n/2 \rfloor$ , then

$$\begin{aligned} \sum_{i=1}^k i 2^{-i} \binom{k}{i} \binom{n-k}{i} &> \sum_{i=1}^{k-1} i 2^{-i} \binom{k}{i} \binom{n-k}{i} \\ &> \sum_{i=1}^{k-1} i 2^{-i} \binom{k-1}{i} \binom{n-k+1}{i}. \end{aligned}$$

Our claim that  $\varphi_s(S) < \varphi_s(S')$  if  $S_\ell$  has fewer leaves than  $S'_\ell$  now follows from (11).  $\square$

## 5.2 Fixed Gene Trees

Suppose now that we are given a fixed gene tree  $T \in R(X)$ ,  $|X| = n \geq 2$ . In this section, we obtain several properties of species trees  $S$  with maximum  $\text{dc}(T, S)$  over all species trees in  $R(X)$ . Although the problem of identifying all species trees  $S$  with maximum  $\text{dc}(T, S)$  for a given gene tree  $T$  remains open for the general case, we solve it for certain classes of cases. Further, we prove some lemmas that apply in the general case.

As in the previous section, we denote by  $S_\ell$  and  $S_r$  the left and right subtrees of the root of  $S$ , and we refer to the sets of leaves of  $S_\ell$  and  $S_r$  as  $X_\ell$  and  $X_r$ .

**Lemma 12.** *Let  $i$  be the number of cherries of  $T$ . Let  $S \in R(X)$  be a species tree with  $k$  leaves in  $S_\ell$  and  $n - k$  leaves in  $S_r$ , where  $k \geq i$  and  $n - k \geq i$ . Then*

$$\text{dc}(T, S) \leq \sigma(T) = \frac{n(n-1)}{2} - i(n-i), \quad (12)$$

with equality if and only if

1. Each cherry of  $T$  is left-right with respect to  $S$ .
2. Both  $S_\ell$  and  $S_r$  are caterpillar trees, and either  $k = i$  or  $n - k = i$  (or both).

**Proof.** By Lemma 4,  $\text{dc}(T, S) \leq m_s(S)$ . Hence, in order to prove (12), it suffices to prove that  $m_s(S) \leq n(n-1)/2 - i(n-i)$ . From (7), we have

$$\begin{aligned} m_s(S) &= -(2n-2) + \sum_{e \in E(S)} n_e \\ &= -(2n-2) + \left( k + \sum_{e \in E(S_\ell)} n_e \right) \\ &\quad + \left( n-k + \sum_{e \in E(S_r)} n_e \right) \\ &= -(n-2) + \sum_{e \in E(S_\ell)} n_e + \sum_{e \in E(S_r)} n_e \\ &\leq -(n-2) + \left( \frac{k(k+1)}{2} - 1 \right) \\ &\quad + \left( \frac{(n-k)(n-k+1)}{2} - 1 \right) \quad (\text{by Lemma 5}) \\ &= \frac{n(n-1)}{2} - k(n-k) \leq \frac{n(n-1)}{2} - i(n-i), \end{aligned}$$

where the last inequality follows from the fact that the function  $-z(n-z)$  is a parabola that faces upward and has line of symmetry  $z = n/2$ .

In order to have equality in (12), we must have

1.  $\text{dc}(T, S) = m_s(S)$ .
2.  $m_s(S) = n(n-1)/2 - i(n-i)$ .

From the derivation,  $m_s(S) = n(n-1)/2 - i(n-i)$  if and only if both  $S_\ell$  and  $S_r$  are caterpillars, and either  $k = i$  or  $n - k = i$  (so that  $-k(n-k)$  is maximum). By Lemma 7,  $\text{dc}(T, S) = m_s(S)$  if and only if each cherry of  $T$  is left-right with respect to  $S$ .  $\square$



































































Denote by  $m_t(T)$  the maximum of  $\text{dc}(T, S)$  across all species trees with the same set of leaves as  $T$ , and denote by  $\varphi_t(T)$  the number of species trees  $S$  that have  $\text{dc}(T, S) = m_t(T)$ .

**Corollary.** *Let  $T$  be a given caterpillar gene tree in  $R(X)$ , where  $|X| = n \geq 2$ . Then  $m_t(T) = \sigma(T) = (n-1)(n-2)/2$ , and  $\varphi_t(T) = (n-1)!$ .*

**Proof.** By Theorem 6,  $(n-1)(n-2)/2$  is an upper bound of  $\text{dc}(T, S)$  for any species tree  $S \in R(X)$ . Thus, we need only show that there are  $(n-1)!$  species trees  $S$  that have



TABLE 2  
The Maximum Deep Coalescence Cost,  $m_t(T)$ , and the Number of Species Trees with Maximum Cost,  $\varphi_t(T)$ , for an Arbitrarily Chosen Labeling of Fixed Gene Tree Shapes with  $1 \leq n \leq 9$  Leaves

$n \leq 5$	 0, 1	 0, 1	 1, 2	 3, 6	 2, 10	 6, 24	 5, 24	 5, 12																				
$n = 6$	 10, 120	 9, 96	 9, 48	 9, 48	 7, 96	 8, 96																						
$n = 7$	 15, 720	 14, 480	 14, 240	 14, 240	 12, 384	 13, 480	 14, 240	 12, 384	 12, 192	 13, 240	 11, 384																	
$n = 8$	 21, 5040	 20, 2880	 20, 1440	 20, 1440	 18, 1920	 19, 2880	 20, 1440	 18, 1920	 18, 960	 19, 1440	 17, 1920	 17, 960	 14, 3072															
$n = 9$	 28, 40320	 27, 20160	 27, 10080	 27, 10080	 25, 11520	 26, 20160	 27, 10080	 25, 11520	 25, 5760	 26, 10080	 24, 11520	 27, 10080	 25, 5760	 22, 11520	 25, 5760	 25, 5760	 25, 5760	 22, 11520	 24, 5760	 25, 5760	 22, 11520	 22, 5760	 24, 2880	 21, 7680	 26, 10080	 24, 5760	 21, 3840	 24, 5760

For each gene tree shape  $T$ , quantities appear as an ordered pair  $(m_t(T), \varphi_t(T))$ . Gene tree shapes are ordered according to their Furnas ranks.

$dc(T, S) = (n - 1)(n - 2)/2$ . Let  $(x, y)$  be the only cherry of  $T$ . The equality conditions in Lemma 12 imply that a species tree  $S$  has this maximum deep coalescence cost if and only if

1.  $S$  is a caterpillar tree.
2. The left subtree  $S_\ell$  is either leaf  $x$  or leaf  $y$ .

If  $|X| = n = 2$ , then  $R(X)$  has just one tree. Hence, there is only one species tree  $S$ , and  $dc(T, S) = 0$ . Suppose that  $|X| = n \geq 3$ . Then we reserve either  $x$  or  $y$  to label the only leaf of  $S_\ell$ , and assign the remaining  $n - 1$  labels to the leaves of  $S_r$ . Because  $S_r$  is also a caterpillar tree, there are  $(n - 1)!/2$  possible labelings of  $S_r$  ([22, Corollary 2.4.3]). Thus, the total number of species trees with  $dc(T, S) = (n - 1)(n - 2)/2$  is  $2 \times (n - 1)!/2 = (n - 1)!$ .  $\square$

Given a caterpillar gene tree  $T$ , the corollary gives a complete set of species trees  $S$  with maximum  $dc(T, S)$ . For a general gene tree  $T$ , obtaining such a complete set is an open problem. However, Lemma 12 implies that  $m_t(T)$  is at least  $\sigma(T)$ . Table 2 lists the values of  $m_t(T)$  and  $\varphi_t(T)$  for all gene tree shapes with up to nine leaves. For gene tree shapes with nine leaves,  $m_t(T)$  and  $\varphi_t(T)$  are also plotted in Fig. 6. Unlike the situation with  $m_s(S)$ , in which  $m_s(S)$  generally decreases as  $\text{rank}_F(S)$  increases, many oscillations occur in  $m_t(T)$  (Fig. 6a). Further,  $m_t(T)$  has relatively little variation across gene tree shapes. For example, the minimum and maximum of  $m_t(T)$  among all 46 gene tree shapes with nine leaves are 21 and 28, whereas in the case of  $m_s(S)$ , the corresponding minimum and maximum are 13 and 28. To compare  $m_t(T)$  with the lower bound  $\sigma(T)$ , Fig. 6a contains a plot of  $\sigma(T)$ . From the figure, we observe that for trees with nine leaves,  $\sigma(T) < \sigma(T')$  implies  $m_t(T) \leq m_t(T')$ . In fact, it can be verified from Table 2 that the relationship holds for any

gene tree shape with up to nine leaves. We conjecture that it is true for arbitrary gene tree shapes: assume that  $T$  has  $i$  cherries and  $T'$  has  $i'$  cherries. Then  $\sigma(T) < \sigma(T')$  implies that  $i > i'$ , that is,  $T$  has more cherries than  $T'$ . The lineages in a species tree  $S$  have fewer chances to coalesce in internal branches of  $S$  to form the cherries of  $T'$  than to form the cherries of  $T$ . Consequently, we expect that more extra lineages are required for reconciling  $T'$  than for reconciling  $T$ .

The plot of  $\varphi_t(T)$  in Fig. 6b contrasts sharply with the plot of  $\varphi_s(S)$  in Fig. 5b. We proved in Section 5.1.3 that  $\varphi_s(S)$  increases as  $\text{rank}_F(S)$  increases. In particular,  $\varphi_s(S)$  has the smallest value (40,320) when  $S$  is a caterpillar tree, and the largest value (567,000) when  $S$  is the most-balanced tree. By contrast,  $\varphi_t(T)$  attains its largest value (40,320) when  $T$  is a caterpillar. Further, this largest value of  $\varphi_t(T)$  is equal to the smallest value of  $\varphi_s(S)$ . Most trees have  $\varphi_t(T) < 15,000$ ; only trees 1, 2, 6, 21, and 23 have  $\varphi_t(T) > 15,000$ . The smallest value of  $\varphi_t(T)$  is 2,880, considerably smaller than the smallest value of  $\varphi_s(S)$ , and it is achieved by four trees: 20, 33, 37, and 38.

We now prove two results that hold for any fixed gene tree  $T$  and provide insight into the features of species trees with maximum  $dc(T, S)$ . We assume in the rest of this section that the gene tree  $T$  has at least two cherries, for otherwise  $T$  is a caterpillar and is covered by the previous corollary. We say that a species tree  $S$  is  $T$ -improvable if there exist two sibling nodes  $q_1$  and  $q_2$  in  $S$  such that  $\{a_1, b_1\} \subseteq C_S(q_1)$  and  $\{a_2, b_2\} \subseteq C_S(q_2)$ , where  $(a_1, b_1)$  and  $(a_2, b_2)$  are two cherries of  $T$  (Fig. 7).

**Lemma 13.** *If  $S$  is  $T$ -improvable, then  $dc(T, S)$  cannot be maximum over all species trees with the same set of leaves as  $S$  and  $T$ .*

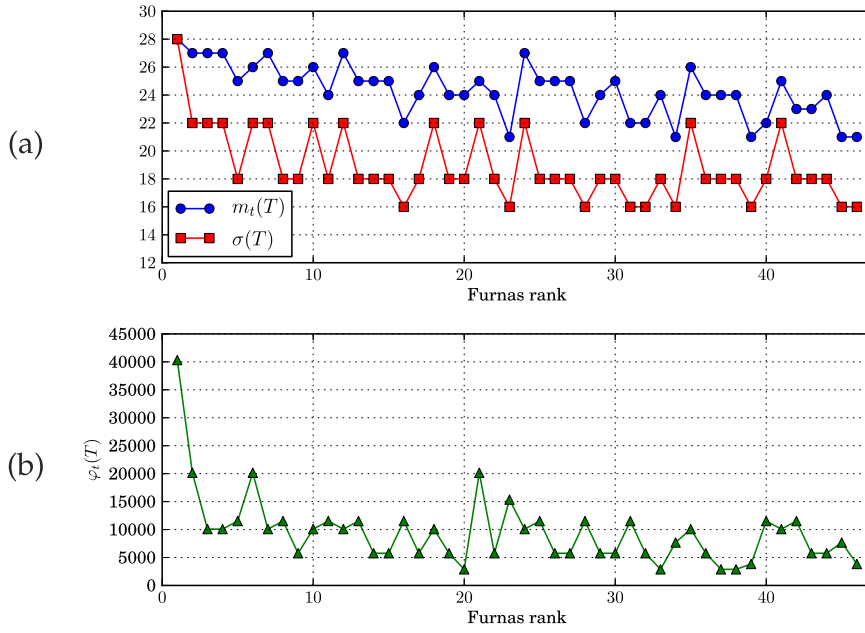


Fig. 6. Plots of  $m_i(T)$ , the maximum deep coalescence cost for fixed  $T$ , and  $\varphi_i(T)$ , the number of species trees  $S$  with  $\text{dc}(T, S) = m_i(T)$ , for all 46 gene tree shapes with nine leaves. The values of  $\sigma(T)$  as lower bounds of  $m_i(T)$  are also plotted. The gene tree shapes are ordered according to their Furnas ranks.

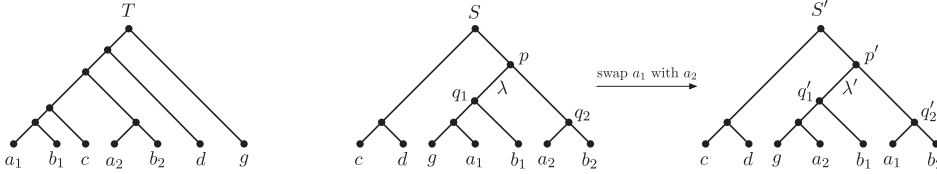


Fig. 7. Illustration for the proof of Lemma 13. In the figure, the gene tree  $T$  has two cherries  $(a_1, b_1)$  and  $(a_2, b_2)$ . In the  $T$ -improvable species tree  $S$ ,  $a_1$  and  $b_1$  are descendant leaves of  $q_1$ , whereas  $a_2$  and  $b_2$  are descendant leaves of  $q_2$ . The species tree  $S'$  is created from  $S$  by swapping leaf  $a_1$  with leaf  $a_2$ . We have  $\text{dc}(T, S') = 7 > \text{dc}(T, S) = 5$ .

To prove the lemma, we explicitly construct a species tree  $S' \neq S$  with  $\text{dc}(T, S') > \text{dc}(T, S)$ . By assumption,  $S$  has two sibling nodes  $q_1$  and  $q_2$  such that  $C_S(q_1)$  contains both leaves of cherry  $(a_1, b_1)$  of  $T$ , and  $C_S(q_2)$  contains both leaves of another cherry  $(a_2, b_2)$  of  $T$ . Let  $S'$  be the tree obtained from  $S$  by swapping leaves  $a_1$  and  $a_2$  (Fig. 7). Two claims are needed for showing that  $\text{dc}(T, S') > \text{dc}(T, S)$ .

**Claim 1.** For each node  $r$  of  $S$ , there exists a unique node  $r'$  of  $S'$  such that exactly one of the following two cases occurs: 1)  $C_S(r) = C_{S'}(r')$ ; 2)  $|C_S(r)| = |C_{S'}(r')|$  and  $(C_S(r) \setminus C_{S'}(r')) \cup (C_{S'}(r') \setminus C_S(r)) = \{a_1, a_2\}$ .

**Proof.** The claim is a direct consequence of the construction of  $S'$  from  $S$ .  $\square$

**Claim 2.** Let  $r$  be a node of  $S$ , and let  $t$  be a maximal subtree of  $T$  with respect to cluster  $C_S(r)$  of  $S$ . If  $t$  contains neither leaf  $a_1$  nor  $a_2$ , then  $t$  is also maximal with respect to cluster  $C_{S'}(r')$  of  $S'$ , where  $r'$  is the unique node of  $S'$  corresponding to  $r$  as determined in Claim 1.

**Proof.** Consider the following cases.

1. Cluster  $C_S(r)$  contains either both  $a_1$  and  $a_2$ , or neither of them. By Claim 1,  $C_{S'}(r') = C_S(r)$ . In this case, any subtree of  $T$  maximal with respect to  $C_S(r)$  is clearly maximal with respect to  $C_{S'}(r')$ .

2. Cluster  $C_S(r)$  contains  $a_1$  but not  $a_2$ . Assume that  $t$  is induced by a node  $v$  of  $T$  (i.e.,  $t = T(v)$ ), and let  $u$  be the parent of  $v$ . We will show that  $C_T(v) \subseteq C_{S'}(r')$  and  $C_T(u) \not\subseteq C_{S'}(r')$ , implying that  $t$  is maximal with respect to  $C_{S'}(r')$ . Because  $t$  is  $C_S(r)$ -maximal,  $C_T(v) \subseteq C_S(r)$ . Along with  $C_{S'}(r') = (C_S(r) \setminus \{a_1\}) \cup \{a_2\}$  (Claim 1) and the assumption that  $t$  does not contain  $a_1$ , this implies  $C_T(v) \subseteq C_{S'}(r')$ .

Assume for a contradiction that  $C_T(u) \subseteq C_{S'}(r')$ . The maximality of  $t$  with respect to  $C_S(r)$  implies that  $C_T(u) \not\subseteq C_S(r)$ . Let  $x$  be an element of  $C_T(u)$  that is not contained in  $C_S(r)$ . If  $x \neq a_2$ , then  $x$  is also not contained in  $C_{S'}(r') = (C_S(r) \setminus \{a_1\}) \cup \{a_2\}$ , implying that  $C_T(u) \not\subseteq C_{S'}(r')$ . Thus,  $x = a_2$  is the only element of  $C_T(u)$  that is not in  $C_S(r)$ . Because  $(a_2, b_2)$  is a cherry of  $T$ , leaf  $b_2$  must be in  $C_T(u)$ , and therefore,  $b_2$  is in  $C_S(r)$ . However,  $C_S(r)$  now contains both leaves  $a_1$  (by assumption) and  $b_2$ , which, respectively, are descendants of sibling nodes  $q_1$  and  $q_2$ . It follows that  $r$  is an ancestor of the parent of  $q_1$  and  $q_2$ . This in turn implies that  $a_2 \in C_S(r)$ , as  $\{a_2, b_2\} \subseteq C_S(q_2)$ , contradicting the assumption that  $C_S(r)$  does not contain  $a_2$ .

3. Cluster  $C_S(r)$  contains  $a_2$  but not  $a_1$ . This case is similar to the preceding case.  $\square$

**Proof of Lemma 13.** For each node  $r$  of  $S$ , let  $r'$  be the unique node of  $S'$  determined in Claim 1. Further, let  $e$  and  $e'$  be the edges of  $S$  and  $S'$  that have  $r$  and  $r'$  as their heads, respectively. We now show case by case that  $\text{xl}(T, e) \leq \text{xl}(T, e')$ , depending on the position of  $r$  relative to the position of node  $p$ , the parent of  $q_1$  and  $q_2$ .

1. Node  $r$  is an ancestor of  $p$ . Then  $C_S(r)$  contains  $C_S(p)$ , which contains both  $a_1$  and  $a_2$ . By Claim 1,  $C_{S'}(r') = C_S(r)$ , and therefore, any  $C_S(r)$ -maximal subtree of  $T$  is also maximal with respect to  $C_{S'}(r')$  and vice versa. By (4),  $\text{xl}(T, e) = \text{xl}(T, e')$ .
2. Node  $r$  is neither an ancestor nor a descendant of  $p$ . Then  $C_S(r)$  contains neither  $a_1$  nor  $a_2$ , and by Claim 1,  $C_{S'}(r') = C_S(r)$ . As in the preceding case, we have  $\text{xl}(T, e) = \text{xl}(T, e')$ .
3. Node  $r$  is a proper descendant of  $p$ . Without loss of generality, we can assume that  $r$  is  $q_1$  or a proper descendant of  $q_1$ . Consider the following subcases.
  - a.  $a_1 \notin C_S(r)$  (e.g.,  $r$  is leaf  $b_1$  in Fig. 7). Clearly,  $a_2 \notin C_S(r)$ , as  $r$  is a descendant of  $q_1$ . By Claim 1,  $C_{S'}(r') = C_S(r)$ , and hence  $\text{xl}(T, e) = \text{xl}(T, e')$ .
  - b.  $a_1 \in C_S(r)$  but  $b_1 \notin C_S(r)$  (e.g.,  $r$  is the node of  $S$  in Fig. 7 that induces cherry  $(g, a_1)$ ). Then  $\{a_1, b_1\} \not\subseteq C_S(r)$ , and by definition, leaf  $a_1$  is maximal with respect to  $C_S(r)$ . We have  $a_2 \in C_{S'}(r') = (C_S(r) \setminus \{a_1\}) \cup \{a_2\}$ , and  $\{a_2, b_2\} \not\subseteq C_{S'}(r')$  (because  $b_2 \notin C_S(r)$ ). Thus, leaf  $a_2$  is a maximal subtree with respect to  $C_{S'}(r')$ . Let  $t$  be a  $C_S(r)$ -maximal subtree of  $T$  other than leaf  $a_1$ . Because  $a_2 \notin C_S(r)$ ,  $t$  does not contain  $a_2$ . By Claim 2,  $t$  is also maximal with respect to  $C_{S'}(r')$ . It follows that  $\text{xl}(T, e) \leq \text{xl}(T, e')$ .
  - c. Both  $a_1$  and  $b_1$  are contained in  $C_S(r)$  (e.g.,  $r$  is node  $q_1$  in Fig. 7). Then  $T$  has a  $C_S(r)$ -maximal subtree  $s$  that contains cherry  $(a_1, b_1)$  of  $T$ . Because  $C_{S'}(r') = (C_S(r) \setminus \{a_1\}) \cup \{a_2\}$ , we have  $\{a_1, b_1\} \not\subseteq C_{S'}(r')$ . Consequently,  $s$  cannot be a  $C_{S'}(r')$ -maximal subtree, and leaves  $a_2$  and  $b_1$  are now  $C_{S'}(r')$ -maximal subtrees. Let  $t$  be a  $C_S(r)$ -maximal subtree of  $T$  other than  $s$ . Clearly  $t$  does not contain  $a_2$  (as  $a_2 \notin C_S(r)$ ), and by Claim 2,  $t$  is maximal with respect to  $C_{S'}(r')$ . Thus, the set of  $C_{S'}(r')$ -maximal subtrees contains every  $C_S(r)$ -maximal subtree, except  $s$ . Further, this set has at least two new subtrees: leaves  $a_2$  and  $b_1$ . By (4),  $\text{xl}(T, e') \geq \text{xl}(T, e) + 1$ .

We have proven that  $\text{xl}(T, e) \leq \text{xl}(T, e')$  for every pair of corresponding edges  $e$  of  $S$  and  $e'$  of  $S'$ . We now show that a pair of edges  $\lambda$  and  $\lambda'$  that falls under case 3c indeed exists. Clearly  $a_1$  and  $b_1$  are in  $C_S(q_1)$ . Hence, for edges  $\lambda = (p, q_1)$  and  $\lambda' = (p', q'_1)$ , we have  $\text{xl}(T, \lambda') > \text{xl}(T, \lambda)$ .  $\square$

The following proposition shows that a species tree that maximizes the cost  $\text{dc}(T, S)$  for a fixed gene tree  $T$  has a certain structure.

**Proposition 14.** Let  $S^*$  be a species tree such that  $\text{dc}(T, S^*)$  is maximum over all species trees in  $R(X)$ . Then

1.  $S^*$  is not  $T$ -improvable.
2. For any two sibling nodes  $q_1$  and  $q_2$ , at least one of  $S^*(q_1)$  and  $S^*(q_2)$  is a caterpillar tree.

**Proof.** Lemma 13 implies the first claim. Then by the definition of a  $T$ -improvable tree, at least one of the pair of subtrees  $S^*(q_1)$  and  $S^*(q_2)$  does not contain both leaves of any cherry of  $T$ . Assume  $S^*(q_1)$  to be that subtree. By (2),

$$\begin{aligned} \text{dc}(T, S^*) &= \sum_{e \in E(S^*)} \text{xl}(T, e) \\ &= \sum_{e \in E(S^*(q_1))} \text{xl}(T, e) \\ &\quad + \sum_{e \notin E(S^*(q_1))} \text{xl}(T, e). \end{aligned} \quad (13)$$

If  $e \notin E(S^*(q_1))$ , then either  $C_{S^*}(e) \supseteq C_{S^*}(q_1)$  or  $C_{S^*}(e) \cap C_{S^*}(q_1) = \emptyset$ . For such an edge  $e$ ,  $\text{xl}(T, e)$ , which can be computed by (4), does not depend on how the leaves of  $S^*(q_1)$  are arranged. It follows that because  $\text{dc}(T, S^*)$  is maximum, the first sum of the right-hand side of (13) must be maximum.

For any internal node  $v$  of  $T$ ,  $\text{MRCA}_{S^*}(v)$  cannot be a node of  $S^*(q_1)$  because  $S^*(q_1)$  does not contain both leaves of any cherry of  $T$ . Hence, for  $e \in E(S^*(q_1))$ ,  $\text{xl}(T, e) = n_e - 1$  by (1). Let  $k$  be the number of descendant leaves of  $q_1$ . Then  $S^*(q_1)$  has  $2k - 2$  edges, and so by Lemma 5, we have

$$\begin{aligned} \sum_{e \in E(S^*(q_1))} \text{xl}(T, e) &= \sum_{e \in E(S^*(q_1))} (n_e - 1) = -(2k - 2) \\ &\quad + \sum_{e \in E(S^*(q_1))} n_e \leq -(2k - 2) \\ &\quad + \left( \frac{k(k+1)}{2} - 1 \right), \end{aligned}$$

with equality in the last equation if and only if  $S^*(q_1)$  is a caterpillar tree. The second claim of the proposition now follows.  $\square$

## 6 DISCUSSION

In this paper, we have studied several properties of the deep coalescence cost. We proved in Section 4 that  $\text{dc}(T, S) = 1$  if and only if  $d_{\text{nni}}(T, S) = 1$ . We also proved that a single rooted-NNI move, applied to either a species tree or a gene tree with  $n$  leaves, can change the deep coalescence cost by at most  $n - 2$ . This result can be useful for branch-and-bound heuristics that search for optimal species trees according to deep coalescence cost by walking through the tree space using rooted-NNI moves. In general, the results in Section 4 can also provide a series of mathematical checks that can assist in verifying the accuracy of deep coalescence algorithms.

The majority of the paper dealt with the problem of determining the maximum deep coalescence cost for either a fixed species tree or a fixed gene tree. As shown in Section 5.1, the maximum deep coalescence cost for a fixed species tree,  $m_s(S)$ , is largest for caterpillar trees, and it generally decreases as the balance of the tree increases. To

a certain extent, this result implies that a balanced species tree needs, on average, a smaller cost to reconcile a gene tree than does a less balanced species tree. We have obtained a formula for the average of  $dc(T, S)$  across all gene trees for a fixed species tree  $S$ , and the plot of the average cost exhibits the same trend observed in the plot of  $m_s(S)$  in Fig. 5a (C.V. Than and N.A. Rosenberg, unpublished data). Thus, our results on  $m_s(S)$  provide an explanation for the tendency of the MDC criterion to infer balanced species trees.

We have completely solved the problem of computing  $m_s(S)$  and  $\varphi_s(S)$ , the number of gene trees with the maximum deep coalescence cost  $m_s(S)$  (Section 5.1). However, the dual problem of computing the maximum deep coalescence cost  $m_t(T)$  for a fixed gene tree remains open. We were able to determine  $m_t(T)$  for caterpillar trees, but we do not have a formula for  $m_t(T)$  in the general case. We have obtained a lower bound for  $m_t(T)$  (Lemma 12), and we have shown some features of species trees  $S$  with  $dc(T, S)$  equal to  $m_t(T)$  (Lemma 13 and Proposition 14).

The present MDC criterion weights the deep coalescence cost  $dc(T, S)$  equally for every gene tree  $T$ . This means that every gene tree  $T$  contributes equally to the total deep coalescence cost, regardless of  $m_t(T)$ . Suppose that for two particular input gene trees  $T_1$  and  $T_2$ ,  $dc(T_1, S) = m_t(T_1)$  and  $dc(T_2, S)$  is equal to, say, half of  $m_t(T_2)$ . Then the choice of  $S$  for reconciling  $T_1$  seems worse than the choice of  $S$  for reconciling  $T_2$ , even if  $m_t(T_2)$  is considerably greater than  $m_t(T_1)$ . The standard MDC criterion does not consider the relationship between a deep coalescence cost and the maximum possible deep coalescence cost for a given input gene tree. We can introduce a normalization by penalizing  $dc(T, S)$  according to  $m_t(T)$ . That is, instead of computing a species tree  $S$  with the minimal total cost  $\sum_{T \in G} dc(T, S)$  for a given collection  $G$  of input gene trees, we could compute a species tree  $S'$  with the minimal sum of normalized costs,  $\sum_{T \in G} dc(T, S')/m_t(T)$ . Such a modified MDC approach could potentially improve upon the inconsistency observed for the standard MDC approach.

## ACKNOWLEDGMENTS

Support was provided by the Burroughs Wellcome Fund and by US National Science Foundation (NSF) grant DBI-1146722.

## REFERENCES

- [1] W.P. Maddison, "Gene Trees in Species Trees," *Systematic Biology*, vol. 46, pp. 523-536, 1997.
- [2] W.P. Maddison and L.L. Knowles, "Inferring Phylogeny Despite Incomplete Lineage Sorting," *Systematic Biology*, vol. 55, pp. 21-30, 2006.
- [3] C. Than and L. Nakhleh, "Species Tree Inference by Minimizing Deep Coalescences," *PLoS Computational Biology*, vol. 5, article e1000501, 2009.
- [4] M.S. Bansal, J.G. Burleigh, and O. Eulenstein, "Efficient Genome-Scale Phylogenetic Analysis under the Duplication-Loss and Deep Coalescence Cost Models," *BMC Bioinformatics*, vol. 11, article S42, 2010.
- [5] H.T. Lin, J.G. Burleigh, and O. Eulenstein, "The Deep Coalescence Consensus Tree Problem is Pareto on Clusters," *Proc. Seventh Int'l Symp. Bioinformatics Research and Applications, Lecture Notes in Computer Science*, vol. 6674, pp. 172-183, 2011.
- [6] L. Zhang, "From Gene Trees to Species Trees II: Species Tree Inference by Minimizing Deep Coalescence Events," *IEEE/ACM Trans. Computational Biology and Bioinformatics*, vol. 8, no. 6, pp. 1685-1691, Nov./Dec. 2011.
- [7] T. Wu and L. Zhang, "Structural Properties of the Reconciliation Space and Their Applications in Enumerating Nearly-Optimal Reconciliations between a Gene Tree and a Species Tree," *BMC Bioinformatics*, vol. 12, article S7, 2012.
- [8] J.H. Degnan and N.A. Rosenberg, "Gene Tree Discordance, Phylogenetic Inference and the Multispecies Coalescent," *Trends in Ecology and Evolution*, vol. 24, pp. 332-340, 2009.
- [9] C.V. Than and N.A. Rosenberg, "Consistency Properties of Species Tree Inference by Minimizing Deep Coalescences," *J. Computational Biology*, vol. 18, pp. 1-15, 2011.
- [10] D.F. Robinson, "Comparison of Labeled Trees with Valency Three," *J. Combinatorial Theory*, vol. 11, pp. 105-119, 1971.
- [11] B. Allen and M. Steel, "Subtree Transfer Operations and Their Induced Metrics On Evolutionary Trees," *Annals of Combinatorics*, vol. 5, pp. 1-13, 2001.
- [12] F.A. Matsen, "A Geometric Approach to Tree Shape Statistics," *Systematic Biology*, vol. 55, pp. 652-661, 2006.
- [13] R. Nichols, "Gene Trees and Species Trees Are Not the Same," *Trends in Ecology and Evolution*, vol. 16, pp. 358-364, 2001.
- [14] J.S. Rogers, "Central Moments and Probability Distributions of three Measures of Phylogenetic Tree Imbalance," *Systematic Biology*, vol. 45, pp. 99-110, 1996.
- [15] M.G.B. Blum and O. Francois, "Minimal Clade Size and External Branch Length under the Neutral Coalescent," *Advances in Applied Probability*, vol. 37, pp. 647-662, 2005.
- [16] R. Klein and D. Wood, "On the Path Length of Binary Trees," *J. Assoc. Computing Machinery*, vol. 36, pp. 280-289, 1989.
- [17] D.E. Knuth, *The Art of Computer Programming: Fundamental Algorithms*, third ed., vol. 1, Addison-Wesley, 1997.
- [18] P.W. Diaconis and S.P. Holmes, "Matchings and Phylogenetic Trees," *Proc. Nat'l Academy of Sciences USA*, vol. 95, pp. 14600-14602, 1998.
- [19] G.W. Furnas, "The Generation of Random, Binary Unordered Trees," *J. Classification*, vol. 1, pp. 187-233, 1984.
- [20] M. Kirkpatrick and M. Slatkin, "Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree," *Evolution*, vol. 47, pp. 1171-1181, 1993.
- [21] C. Semple and M. Steel, *Phylogenetics*, Oxford Lecture Series in Mathematics and Its Applications, vol. 24, Oxford Univ. Press, 2003.



**Cuong V. Than** received the PhD degree in computer science from Rice University in 2009. He was at the University of Michigan from 2009 to 2011 for postdoctoral training, and he is currently a postdoctoral researcher at Stanford University under the supervision of Professor Noah Rosenberg. His main research interests include phylogenetics.



**Noah A. Rosenberg** received the PhD degree in biological sciences from Stanford University in 2001 and completed postdoctoral training at the University of Southern California. From 2005 to 2011, he served as a faculty member at the University of Michigan, and he is now associate professor in the Department of Biology at Stanford University. His research interest focuses on human evolutionary genetics, population-genetic theory, and mathematical phylogenetics.

► For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/publications/dlib](http://www.computer.org/publications/dlib).